Project brief

# Overview

Develop approaches for cost curve optimization for ERW - accounting for the spatial heterogeneity and measurement costs in the analysis

# Strategy

(Code : <https://github.com/PSU-Geofluids-Lab/Geospatial_Analysis.git> )

### Approach

1. Generate spatial models - either from data or from synthetic realizations like chosen variograms – analyze how soil properties change spatially (variograms, variogram maps, VQ-VAE ? - tokenization)
   1. Can use ISDA (critical to remind that this is NOT generated from variogram or spatial statistics explicitly) - https://www.isda-africa.com/isdasoil/technical-information/
      1. Could just use ISDA as is and say “This is the sample”
      2. Could use the ISDA variance estimate to generate a sample at each point with a given mean and variance
         1. This would really only be for a “worked example”
         2. No really accounting for spatial correlations in variance
            1. Pixel by pixel framework kind of glosses over this anyways
   2. Can make our own variograms and/or spatial generative models
      1. Like for spatial generative: <https://porespy.org/_examples/generators/index.html>
      2. Fractal noise at different scales – exemplar of what it looks like in variogram space/variogram maps
   3. Can ISDA actual point measurements of the soil: this would become most ERW practitioner's database equivalent. <https://www.isda-africa.com/isdasoil/technical-information/> - https://github.com/iSDA-Africa/open-soil-data/

Jake’s Proposed plan of attack: Compare results from ii and iii. This has a few levels of utility. We start with something synthetic where we know the real answers to everything but are looking at a map that is intrinsically generative (i.e. we aren’t directly controlling features but it can be reproducible). Then we look at a public analogue of data that practitioners can imagine themselves using. Then we say: hey is it important to make a variogram in the first place?

VQ-VAE: One pathway down the road after the variogram (inevitably?) does not work. [Probably best to just mention this in passing and then move along]

Overall structure :

Synthetic - > Real data -> Complex generative model (ISDA)

**End goal : A map over a chosen area that you can sample from for the next step.**

1. Use the variability to sample for the end-members you would use for soil mixing – and then after choosing a basalt end-member (with variance) and a specific amount of weathering, bootstrap resample to get a range of predictions for the weathering amount – get the PDF and see how many measurements you need to get a required level of precision in the estimated weathering and the associated cost (get the accuracy vs cost estimate) (conceptually there is some similarity to this re fertilizer usage and optimization cost curve, max profit given a total time/effort ..)

Basalt endmember has “user-defined” distribution (this is basically a bunch of ICP-MS measurements that were taken from a bunch of hand samples: [we can play with the uncertainty in this as well…+-10%? Etc etc - eg., XRF vs ICP etc ]).

Digital Twin? To do a full digital twin you would need a forward model : here the “mixing model” is a version of the forward model to generate the synthetic data, though one could use a reactive transport model of varying accuracy to do an analysis also (of varying levels of complexity).

Practically, the idea would be the following :

1. Sample from the spatial map of variance -> then do the same for the basalt data (sampling from known points, and or expectations of variation in the composition) → Then use one of the soil samples as the mixing end-member (for one post-deploymenet sample, can be different for each sampling or assign regularization re re spatial location of sampling etc), and the other sample is used to mix basalt/rock (chosen amount) + chosen dissolution fraction → this generate the synthetic mixing data (like what you may measure) + add extra uncertainty due to method error ..

We assign a cost to each measurement and look at how the convergence of the mean of the estimated dissolution works out. The model for the first pass case is v simple since its a mixing model, but we can make it more complex and/or add multiple element-based inversion and/or alkalinity and/or pH and/or mineralogical measurements.

We are sort of merging inverse model part here to build cost curves and optimization while there are aspects of classical Digital Twin with a generative process model and intervention/strategy testing and some data assimilation ..

1. Start with a three end-member model like TiCat (<https://cdrxiv.org/preprint/390> - Eqn 7) https://porespy.org/\_examples/generators/index.htmland later consider the benefit of having a more realistic weathering model with specific minerals (so more end-member mixing, more params to fit/degeneracy - use MCMC for the inversion). Can include things like alkalinity and/or pH in the mix also (maybe a simpler version of this ? <https://github.com/MatteoBertagni/SMEW>)

**Key Goals :**

1. **Cost Curve!!!** (and finding the optimum for it in terms of measurements)
2. **Location Optimization:** Considering optimization of site selection for new locations in the context of available data and expected variance. A key point is that with higher variance, it becomes more expensive to measure the small amount of dissolution – so, there is a tradeoff re detectibility of dissolution, tradeoff for costs to measure the signal in the chosen data, and the return on investment needed to make this happen.

**General Data/Analysis steps :**

**Hypothesis to test :**

**A** What is the relative role of mineralogy vs elemental data alone for ERW calcs. We can get mineralogy data for the basalt samples going in with xrd and some spectroscopy methods, and even to first order from CIPW norm type calcs. And also presumably getting some measurements of the mineralogy in soil (esp plag, olivine, cpx ratios) could be a better discriminators for weathering estimates than many other complex trace element measurements.

A while back, I had pulled a big basalt compilation from georoc and done cipw to get normative mineralogy (it’s not perfect, but based on spot checks it gave decent results). So maybe a good next steps.

* doing some inversion tests with the forward model data to look at the effects re uniqueness, data noise, most relevant variables (this can be layered on top of the base spatial analysis parts).
* do this type of exercise for lots of rock data - by automating the process and seeing if there are some general trends vs variation across locations and composition.
* can add a simple layer re relative reactivity of different minerals - I know that there are strong limitations with the kinetics data in terms of ph - eH - T space coverage, but we can start with the big compilations from Sue and others from USGS to make some informed choices. Then rinse and repeat part b, but with some physically bounded constraints re relative fraction of weathering and over time.